Unsupervised individual whales identification: spot the difference in the ocean

Alexis Joly[1], Jean-Christophe Lombardo[1], Julien Champ[1], and Anjara Saloma[3]

[1] Inria ZENITH team, LIRMM, France, name.surname@inria.fr
[2] INA, France, obuisson@ina.fr
[3] Cetamada, NGO, anjara@cetamada.com

Abstract. Identifying organisms is a key step in accessing information related to the ecology of species. But unfortunately, this is difficult to achieve due to the level of expertise necessary to correctly identify and record living organisms. To try bridging this gap, enormous work has been done on the development of automated species identification tools such as image-based plant identification or audio recordings-based bird identification. Yet, for some groups, it is preferable to monitor the organisms at the individual level rather than at the species level. The automatizing of this problem has received much less attention than species identification. In this paper, we address the specific scenario of discovering humpack whales individuals in a large collections of pictures collected by nature observers. The process is initiated from scratch, without any knowledge on the number of individuals and without any training samples of these individuals. Thus, the problem is entirely unsupervised. To address it, we set up and experimented a scalable fine-grained matching system allowing to discover small rigid visual patterns in highly clutter background. The evaluation was conducted in blind in the context of the LifeCLEF evaluation campaign. Results show that the proposed system provides very promising results with regard to the difficulty of the task but that there is still room for improvements to reach higher recall and precision in the future.

1 Introduction

Identifying organisms is a key step in accessing information related to the ecology of species. This is an essential step in recording any specimen on earth to be used in ecological studies. But unfortunately, this is difficult to achieve due to the level of expertise necessary to correctly identify and record living organisms. Watson et al.[6] discussed in 2004 the potential of automated species identification approaches typically based on machine learning and multimedia data analysis methods. They suggested that, if the scientific community is able to (i) overcome the production of large training datasets, (ii) more precisely identify and evaluate the error rates, (iii) scale up automated approaches, and (iv) detect novel species, it will then be possible to initiate the development of a generic automated species identification system that could open up vistas of new opportunities for pure and applied work in biological and related fields. Since the question raised by Watson in 2004 (”automated species identification: why not?”), enormous work has been done on the development of effective methods such as image-based plant identification [2,8,7], bird songs identification [22], fish species identification [20], etc.

The problem of automatically identifying individual organisms rather than species has received much less attention (except for humans of course). Yet, for some groups, it is preferable to monitor the organisms at the individual level rather than at the species level. This is notably the case of big animals, such as whales and elephants, whose population are scarcer and who are traveling longer distances. Monitoring individual animals allow gathering valuable information about population sizes, migration, health, sexual maturity and behavior patterns. Tracking devices and tagging technologies are only part of the solution because of their invasive character, relatively high cost and limited lifetime. Morphological/biometric approaches are a complementary approach that is less invasive, more durable and cheaper for nature observers mobilized on a given spot. Using natural markings to identify individual animals over time is usually known as *photo-identification*. This research technique is used on many species of marine mammals. Initially, scientists used artificial tags to identify individual whales, but with limited success (most tagged whales were actually lost or died). In the 1970s, scientists discovered that individuals of many species could be recognized by their natural markings. These scientists began taking photographs of individual animals and comparing these photos against each other to identify individual animal’s movements and behavior over time. Since its development, photo-identification has proven to be a useful tool for learning about many marine mammal species including humpbacks, right whales, finbacks, killer whales, sperm whales, bottlenose dolphins and other species to a lesser degree. Nowadays, this process is still mostly done manually making it impossible to get an accurate count of all the individuals in a given large collection of observations. Researchers usually survey a portion of the population, and then use statistical formulae to determine population estimates. To limit the variance and bias of such an estimator, it is however required to use large-enough samples which still makes it a very time-consuming process. Automating the *photo-identification* process could drastically scale-up such surveys and open brave new research opportunities for the future.

In this paper, we address more particularly the problem of discovering all individual humpack whales appearing in a large collection of caudal’s images in a fully unsupervised way, *i.e.* without any knowledge on the number of individuals and without any training samples of these individuals. This is in essence a different and more challenging problem than the supervised recognition of individual whales such as the challenge proposed by NOAA Fisheries through the Kaggle platform[4]. Such supervised scenario is actually only affordable when the individuals are already well known and well illustrated by tens of pictures that were hardly collected along the years. On the other side, the unsupervised identifica-

[4] `https://www.kaggle.com/c/noaa-right-whale-recognition`

tion scenario targeted in this paper has the great advantage to allow the use of unlabeled or very labeled collections of observations which is the vast majority of available data today.

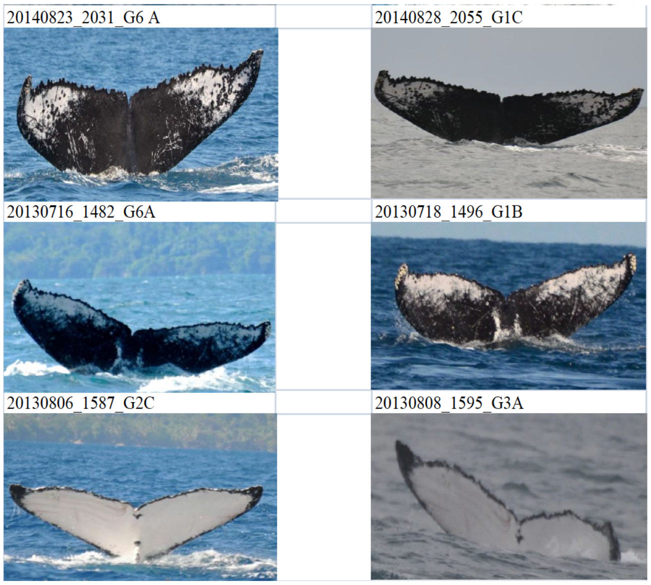


Fig. 1. Three good matches (each line corresponds to 2 images of the same individual whale)

2 Data and challenge

The experiment reported in this paper was part of the 2016-th edition of the the LifeCLEF international evaluation campaign [14] (in particular in the scope of the sea organisms identification task). The data shared through this challenge consisted of 2005 images of humbpack whales caudals collected by the CetaMada[5] NGO between 2009 and 2014 in Madagascar area. Cetamada is a Malagasy Non-Profit Association created in May, 2009 whose goal is to protect marine mammal population and their habitat in Madagascar through sustainable eco-tourism and scientific research. There are presently 4 citizen sciences data collection sites (St. Marys, Majunga, Ifaty and Fort Dauphin) for which hotel-establishements and their customers have become sentinels for data collection. This method helps obtain more than 250 photo IDs each year, which effectively helps produce a photo catalogue of humpback whales reproducing on Malagasy coasts.
After acquisition, each photograph was manually cropped so as to focus only on

[5] `https://www.cetamada.org/`

the caudal fin that is the most discriminant pattern for distinguishing an individual whale from another. Figure 1 displays six of such cropped images, each line corresponding to two images of the same individual. As one can see, the individual whales can be distinguished thanks to their natural markings and/or the scars that appear along the years. Automatically finding such matches in the whole dataset and rejecting the false alarms is difficult for three main reasons. The first reason is that the number of individuals in the dataset is high, around $1,200$, so that the proportion of true matches is actually very low (around 0.05% of the total number of potential matches). The second difficulty is that distinct individuals can be very similar at a first glance as illustrated by the false positive examples displayed in Figure 2. To discriminate the true matches from such false positives, it is required to detect very small and fine-grained visual variations such as in a spot-the-difference game. The third difficulty is that all images have a similar water background of which the texture generates quantities of local mismatches.

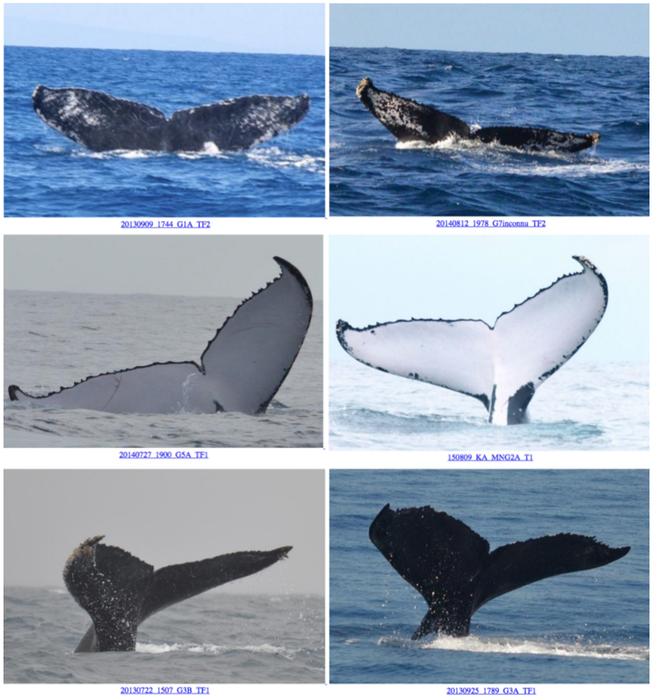


Fig. 2. Three false positives (each line corresponds to 2 distinct individual whales)

3 Method Description

To differentiate biomarkers from the mass of other visual patterns without any supervision, the research line we investigate in this work is to rely on the spatial filtering of low-level visual correspondences. Our hypothesis is that the biomarkers are sufficiently localized on the fin to be considered as non deformable objects so that two views of the same biomarker in two different images are supposed to be related by epipolar geometry. On the other side, the raw noisy visual correspondences at the origin of false alarms should be filtered by the use of geometric rules. The standard solution to perform such epipolar geometry estimation is to use the RANSAC algorithm [5]; it consists in generating transformation hypotheses using a minimal number of low-level visual correspondences and then evaluating each hypothesis based on the number of *inliers* among all features under that hypothesis. The main advantage of the RANSAC algorithm is that it is robust to the presence of a high number of *outliers* which makes it suitable to deal with the large numbers of false alarms that are generated by the raw visual matching of the local features.

As the RANSAC algorithm can be rather slow, an efficient variant, LO-RANSAC, was proposed by Chum et al. [3] and has been proved to provide consistent speed-ups in many image retrieval frameworks [19,18,17,1]. It involves generating hypotheses of an approximate model thanks to the shape information provided with the affine-invariant image regions from which the visual features were extracted. With this method, an hypothesis can be generated with only a single pair of corresponding features whereas two or three are required when using only the feature positions. This greatly reduces the number of possible hypotheses which need to be considered by the RANSAC algorithm and significantly speeds up the spatial verification procedure. An even faster strategy [10], consists in considering only the shape information of the image regions, without exploiting the positions of the features at all. A rough approximation of the best transformation can then actually be estimated by a Hough-like voting strategy on the quantized differences of the characteristic orientation and scale of each visual correspondence. Using this so-called *weak geometry* method allows trading quality for time and is the only acceptable solution when dealing with huge image sets and real-time contexts (e.g. a search engine working on billions of images).

The spatial verification we use in our own system is also a variant of the RANSAC algorithm making use of weak geometry rules generated from the region shape characteristics. We however do not use the weak geometry to directly generate an hypothesis from a single visual correspondence. We rather use it to filter the exact hypothesis generated by the classical RANSAC algorithm. Concretely, if we restrict our class of transformations to rotation and scaling, the RANSAC algorithm can generate an hypothesis from any pair of visual correspondences. To quickly decide whether this hypothesis is relevant or not, we check its consistency with regard to the two approximate hypothesis generated from the shape characteristics of each visual correspondence. If any of the two approximate models does not fit the RANSAC hypothesis, we reject that solution without computing the costly consensus phase. In practice, up to 99% of the RANSAC hypothesis can be rejected in that way (leading to a consistent speed-up).

Another major difference between our method and the ones in [19,18,17,1] is that we use the ranking of the visual correspondences to further improve the matching. Our retrieval framework does actually not rely on the popular bag-of-words model to generate the raw visual correspondences but on a more accurate approximate KNN search algorithm (described in section 4). The main benefit is that the precision of our raw visual matches is already much better than the ones produced by the bag-of-words model (based on vector quantization). The RANSAC algorithm therefore works on less correspondences and less false alarms. Another benefit is that each raw visual correspondence $\{\mathbf{x}, \mathbf{y}\}$ is associated with a rank $r_{\mathbf{x}}(\mathbf{y})$). This allows two things: (i) to restrict the generation of the hypothesis of the RANSAC algorithm to the best match of each feature $\mathbf{x}$ in the transformed image I_Y. The number of evaluated hypothesis is consequently reduced, particularly in the presence of numerous repeated visual patterns (the burstiness phenomenon [9]) (ii) the ranking can be used in the computation of the final score by weighing the contribution of each inlier according to its rank in the whole dataset. Closest points are then favored to the detriment of the farthest ones, independently from the feature space density in the neighborhood of x_q. More formally, given a couple of images I_X and I_Y, represented by sets of local features X and Y, we define the following spatially consistent match kernel:

$$K(I_X, I_Y) = \frac{1}{2}\left(S_X(I_Y) + S_Y(I_X)\right) \tag{1}$$

$$S_X(I_Y) = \sum_{\mathbf{x} \in X} \max_{\mathbf{y} \in Y} \left[\delta_{X,Y}(\mathbf{x}, \mathbf{y}).\varphi(r_{\mathbf{x}}(\mathbf{y}))\right] \tag{2}$$

where $r_{\mathbf{x}}(\mathbf{y}) : \mathbb{R}^d \rightarrow \mathbb{N}^+$ is a ranking function that returns the rank of the local feature $\mathbf{y}$ according to its L_2-distance to the query feature $\mathbf{x}$ (within the whole dataset). The function $\varphi()$ is a decreasing function allowing to give more weights to the top ranked features (we used the inverse function in our experiments). Finally, $\delta_{X,Y}(\mathbf{x}, \mathbf{y})$ is an indicator function equal to one if the correspondence $(\mathbf{x}, \mathbf{y})$ is an inlier of the geometric model estimating the transformation between I_X and I_Y, *i.e.*:

$$\delta_{X,Y}(\mathbf{x}, \mathbf{y}) = \left(\left\|\mathbf{P}_{\mathbf{x}} - (\hat{\mathbf{A}}\mathbf{P}_{\mathbf{y}} + \hat{\mathbf{B}})\right\| < \theta\right) \tag{3}$$

where $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ are the parameters of the best transformation estimated by our accelerated RANSAC algorithm, $\mathbf{P}_{\mathbf{x}}$ and $\mathbf{P}_{\mathbf{y}}$ are the spatial positions of $\mathbf{x}$ and $\mathbf{y}$, θ is a user defined threshold defining the spatial tolerance of the inliers (we used $\theta = 16$ pixels in our experiments). The number of probes of the RANSAC algorithm was set to 10K in all our experiments.

4 Approximate K-NN search scheme

In practice, to speed up the computation of the matching, the ranking function $r_{\mathbf{x}}(\mathbf{y}) : \mathbb{R}^d \rightarrow \mathbb{N}^+$ is implemented as an approximate nearest neighbors search algorithm based on hashing and probabilistic accesses in the hash table. It takes as input a query feature $\mathbf{x}$ and an inverted index of all local features $\mathbf{z} \in \mathcal{Z}$ extracted from the image collection. It returns a set of m approximated neighbors with an approximated rank $r_{\mathbf{x}}^{m}(\mathbf{y})$ (we used $m = 500$ in all our experiments). The exact ranking function $r_{\mathbf{x}}(\mathbf{y})$ is simply replaced by this approximated ranking function in all equations above. Note that the features $\mathbf{y}$ that are not returned in the top-m approximated nearest neighbors are simply *removed* from the match kernel equations conducting to a considerable reduction of the computation time. Consequently, they are implicitly considered as having a rank-based activation function $\varphi(r_{\mathbf{x}}(\mathbf{y}))$ equal to zero which is a good approximation as their rank is supposed to be higher than m. The higher the value of m is and the lower the error compared to the exact match kernel.

Let us now describe more precisely our approximate nearest neighbors indexing and search method. It first compresses the original feature vectors $\mathbf{z} \in \mathcal{Z}$ into compact binary hash codes $\mathbf{h}(\mathbf{z})$ of length b thanks to the use of a data-dependent high-dimensional hash function. In our experiments, we used RMMH hash function [13] that has the advantage to be easily implemented and to be effective for any kind of visual features or data distribution. The distance between any two features $\mathbf{x}$ and $\mathbf{z}$ can then be efficiently approximated by the Hamming distance between $\mathbf{h}(\mathbf{x})$ and $\mathbf{h}(\mathbf{y})$. In all our experiments we used $b = 128$ bits.

To avoid scanning the whole dataset, the hash codes $\mathbf{h}(\mathbf{z})$ derived from the local features of the entire set $\mathcal{Z}$ are then indexed in a hash table whose keys are the t-length prefix of the hash codes $\mathbf{h}(\mathbf{z})$. At search time, the hash code $\mathbf{h}(\mathbf{x})$ of a query feature $\mathbf{x}$ is computed as well as its t-length prefix. We then use a probabilistic multi-probe search algorithm inspired by the one of [11] to select the buckets of the hash table that are the most likely to contain exact nearest neighbors. This is done by using a probabilistic search model that is trained offline on the exact m-nearest neighbors of M sampled features $\mathbf{z} \in \mathcal{Z}$. We however use a simpler search model than the one of [11]. We actually use a normal distribution with independent components parameterized by a single vector σ that is trained over the exact nearest neighbors of the training samples. At search time, we also use a slightly different probabilistic multi-probe algorithm trading stability for time. Instead of probing the buckets by decreasing probabilities, we rather use a greedy algorithm that computes the probability of neighboring buckets and select only the ones having a probability greater than a threshold ζ that is fixed over all queries. The value of ζ is trained offline on M training samples and their exact nearest neighbors so as to reach on average cumulative probability α over the visited buckets. In our experiments, we always used $\alpha = 0.95$ meaning that on average we retrieve 95% of the exact nearest neighbors in the original feature space. Once the most probable buckets have been selected, the refinement step computes the Hamming distance between $\mathbf{h}(\mathbf{x})$ and the $\mathbf{h}(\mathbf{z})$'s belonging to the selected buckets and keep only the top-m matches thanks to a max heap.

5 Experiments

As mentioned earlier, the system described above was evaluated in the context of the Sea task of the LifeCLEF 2016 evaluation campaign [14]. This means that the experiment was conducted in blind, *i.e.* without having access to the ground truth, and thus, without any possibility of learning or tuning the parameters of the system.

5.1 Task Description

The task was simply to detect as many true matches as possible from the whole dataset, in a fully unsupervised way. Each evaluated system had to return a *run file* (i.e., a raw text file) containing as much lines as the number of discovered matches, each match being a triplet of the form:

$$< imageX.jpg\ imageY.jpg\ score >$$

where *score* is a confidence score in $[0, 1]$ (1 for highly confident matches). The retrieved matches had to be sorted by decreasing confidence score. A run should not contain any duplicate match (e.g., $< image1.jpg\ image2.jpg\ score >$ and $< image2.jpg\ image1.jpg\ score >$ should not appear in the same run file). The metric used to evaluate each run is the Average Precision:

$$AveP = \frac{\sum_{k=1}^{K} P(k) \times rel(k)}{M}$$

where M is the total number of true matches in the groundtruth, k is the rank in the sequence of returned matches, K is the number of retrieved matches, $P(k)$ is the precision at cut-off k in the list, and $rel(k)$ is an indicator function equaling 1 if the match at rank k is a relevant match, 0 otherwise. The average is over all true matches and the true matches not retrieved get a precision score of 0.

5.2 Submitted runs

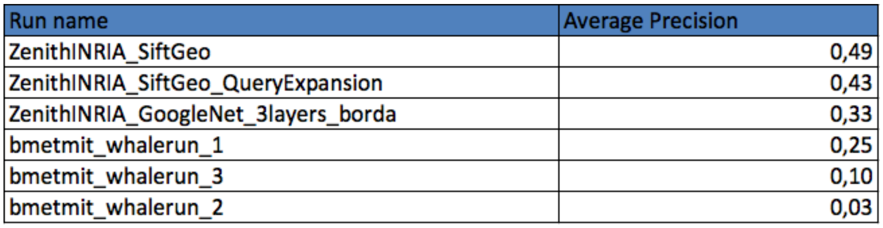
We submitted a total of 3 *run files* to the LifeCLEF benchmark corresponding to three configurations of our system. Each run was computed by (i) searching each image of the collection one by one, (ii) computing the score K of the image pairs according to Equation 1 and (iii) rank all pairs by decreasing value of K.

- Run *ZenithINRIA_SiftGeo*: In this run we used SIFT local features [15] extracted around Harris Hessian regions [16] (without threshold).
- Run *ZenithINRIA_GoogleNet_3layers_borda*: In this run we used off-the-shelf local features extracted at three different layers of GoogLeNet convolutional neural network [21] (layer *conv2-3x3*: 3136 local features per image, layer *inception_3b_output*: 784 local features par image, layer *inception_4c_output*: 196 local features per image). The matches found using the 3 distinct layers were merged through a late-fusion approach based on Borda.

- Run *ZenithINRIA_SiftGeo_QueryExpansion*: This the last run differs from the run *ZenithINRIA_SiftGeo* in that a query expansion strategy was used to re-issue the regions matched with a sufficient degree of confidence as new queries (using the method described in [12]).

5.3 Official LifeCLEF results

The table in Figure 3 provides the scores achieved by the three configurations of our system as well as the scores obtained by the system of the other competitor and described in [4] (using a Fisher Vector image representation based on SIFT features and a GMM visual codebook of 256 visual words). Runs *bmetmit_whalerun_2* and *bmetmit_whalerun_3* differ from *bmetmit_whalerun_1* in that segmentation propagation was used beforehand so as to separate the background (the water) from the whales caudal fin.



Run name	Average Precision
ZenithINRIA_SiftGeo	0,49
ZenithINRIA_SiftGeo_QueryExpansion	0,43
ZenithINRIA_GoogleNet_3layers_borda	0,33
bmetmit_whalerun_1	0,25
bmetmit_whalerun_3	0,10
bmetmit_whalerun_2	0,03

Fig. 3. Individual whale identification results: AP of the 6 evaluated systems

The main conclusion we can draw from the results of this evaluation is that the spatial arrangement of the local features is a crucial information for rejecting the false positives (as proved by the much higher Average Precision of our system compared to the one of *bme_mit*). As powerful as aggregation-based methods such as Fisher Vectors are for fine-grained classification, they do not capture the spatial arrangement of the local features which is a precious information for rejecting the mismatches without supervision. Another reason explaining the good performance of the best run *ZenithINRIA_SiftGeo* is that it is based on affine invariant local features contrary to *ZenithINRIA_GoogleNet_3layers_borda* and *bme_mit* runs that use grid-based local features. Such features are more sensitive to small shifts and local affine deformations even when learned through a powerful CNN such in our run *ZenithINRIA_GoogleNet_3layers_borda*. The comparison of our two runs *ZenithINRIA_SiftGeo* and *ZenithINRIA_SiftGeo_QueryExpansion* show that query expansion did not succeeded in improving the results. Query expansion is actually a risky solution in that it is highly sensitive to the decision threshold used for selecting the re-issued matched regions. It can be considerably increase recall when the decision threshold is well estimated but at the opposite, it can also boost the false positives when the threshold is too low.

5.4 Additional results

To get a more practical understanding of the performance achieved by our system, Figure 4 plots the recall-precision curve of our best run. It shows that the main strength of our system is that it has a very high precision on the best found matches (thanks to the spatial filtering). Actually, the top-100 matches were found with a perfect precision of 1.00 which makes our system already usable for automatically discovering some matches without any human control. On the other side, our system fails in reaching high recall values automatically. It would require further human validation to reach reasonable recalls in the range of $30-80\%$. But still, this would be a much more easier process than discovering the matches from scratch.

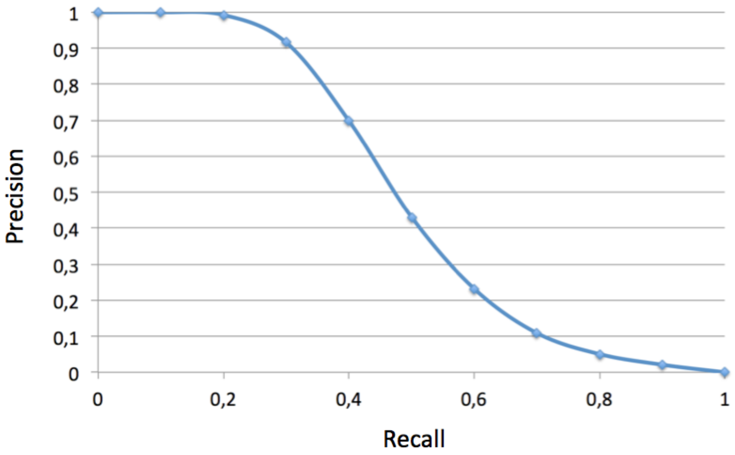


Fig. 4. Recall-Precision curve of our best system (Affine SIFT + Geometry)

Table 1 provides the search processing time of each run. It shows that using the Affine SIFT features or the off-the-shelf CNN features requires an equivalent amount of time. The total search time for discovering all the matches in the image collection was about 24 hours. This is yet not negligible but definitely acceptable compared to the difficulty of doing that manually. One should also notice that we used a very high quality approximate nearest neighbors search ($alpha$ = 95%) to favor quality over time. Much more faster runs could be obtained using moderate values of $alpha$ (*e.g.* 80%) without degrading much the results.

Table 1: Search time of three configurations of our system

Run name	**Total time**	Avg time per query image
ZenithINRIA_SiftGeo	86 415s	43.1s
ZenithINRIA_GoogleNet_3layers_borda	67 969s	33.9s
ZenithINRIA_SiftGeo_QueryExpansion	31 494s	119.0s

6 Conclusion

In this paper, we addressed the problem of identifying humpack whales individuals in a large collections of aerial pictures of caudal fins in a fully unsupervised way. We therefore designed a scalable fine-grained matching system allowing to discover small rigid visual patterns in highly clutter background. It was experimented in the context of a blind system-oriented evaluation in which it performed the best. The comparison to the other evaluated system show that the spatial arrangement of the local features is a crucial information to discriminate the individual whales as well as to filter the potentially huge number of false positive matches. Overall, the Average Precision of our system is about 49. This is still not satisfactory for a fully automatic detection scenario but, on the other side, this might already drastically simplify the manual work of the biologists through the release of interactive validation tools. In further work, we will attempt to use localized spatially consistent similarities rather than estimating a global affine transformation at the image level. Also, we will explore possible extensions of convolutional auto-encoders as a way to discover the semi-deformable bio-markers.

References

1. Arandjelovic, R., Zisserman, A.: Three things everyone should know to improve object retrieval. In: Proc. CVPR (2012)
2. Bruno, O.M., de Oliveira Plotze, R., Falvo, M., de Castro, M.: Fractal dimension applied to plant identification. Information Sciences 178(12), 2722–2733 (2008)
3. Chum, O., Matas, J., Obdrzalek, S.: Enhancing ransac by generalized model optimization. In: Proc. of the ACCV. vol. 2, pp. 812–817 (2004)
4. Dävid Papp, D.L., Szücs, G.: Object detection, classification, tracking and individual recognition for sea images and videos. In: Working notes of CLEF (2016)
5. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24(6), 381–395 (1981)
6. Gaston, K.J., O'Neill, M.A.: Automated species identification: why not? Philosophical Transactions of the Royal Society of London B: Biological Sciences 359(1444), 655–667 (2004)
7. Goëau, H., Bonnet, P., Joly, A., Bakic, V., Barthélémy, D., Boujemaa, N., Molino, J.F.: The imageclef 2013 plant identification task. In: CLEF (2013)
8. Hossain, J., Amin, M.A.: Leaf shape identification based plant biometrics. In: Computer and Information Technology (ICCIT), 2010 13th International Conference on. pp. 458–463. IEEE (2010)

9. Jégou, H., Douze, M., Schmid, C.: On the burstiness of visual elements. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 1169–1176. IEEE (2009)
10. Jégou, H., Douze, M., Schmid, C.: Improving bag-of-features for large scale image search. International Journal of Computer Vision 87(3), 316–336 (2010)
11. Joly, A., Buisson, O.: A posteriori multi-probe locality sensitive hashing. In: Proceedings of the 16th ACM international conference on Multimedia. pp. 209–218. ACM (2008)
12. Joly, A., Buisson, O.: Logo retrieval with a contrario visual query expansion. In: Proceedings of the 17th ACM international conference on Multimedia. pp. 581–584. ACM (2009)
13. Joly, A., Buisson, O.: Random maximum margin hashing. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. pp. 873–880. IEEE (2011)
14. Joly, Alexis and Goëau, Hervé and Glotin, Hervé and Spampinato, Concetto and Bonnet, Pierre and Vellinga , Willem-Pier and Champ, Julien and Planqué, Robert and Palazzo, Simone and Müller, Henning booktitle=Proceedings of CLEF 2016, y.: Lifeclef 2016: multimedia life species identification challenges
15. Lowe, D.G.: Object recognition from local scale-invariant features. In: Computer vision, 1999. The proceedings of the seventh IEEE international conference on. vol. 2, pp. 1150–1157. Ieee (1999)
16. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. International journal of computer vision 60(1), 63–86 (2004)
17. Perd'och, M., Chum, O., Matas, J.: Efficient representation of local geometry for large scale object retrieval. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 9–16. IEEE (2009)
18. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2008)
19. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. pp. 1–8. IEEE (2007)
20. Shortis, M.R., Ravanbakskh, M., Shaifat, F., Harvey, E.S., Mian, A., Seager, J.W., Culverhouse, P.F., Cline, D.E., Edgington, D.R.: A review of techniques for the identification and measurement of fish in underwater stereo-video image sequences. In: SPIE Optical Metrology 2013. pp. 87910G–87910G. International Society for Optics and Photonics (2013)
21. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9 (2015)
22. Tyagi, H., Hegde, R.M., Murthy, H.A., Prabhakar, A.: Automatic identification of bird calls using spectral ensemble average voice prints. In: Signal Processing Conference, 2006 14th European. pp. 1–5. IEEE (2006)